Identifiability of 2-tree mixtures for group-based models

Elizabeth S. Allman, Sonja Petrović, John A. Rhodes, and Seth Sullivant

Abstract—Phylogenetic data arising on two possibly different tree topologies might be mixed through several biological mechanisms, including incomplete lineage sorting or horizontal gene transfer in the case of different topologies, or simply different substitution processes on characters in the case of the same topology. Recent work on a 2-state symmetric model of character change showed that for 4 taxa such a mixture model has non-identifiable parameters, and thus it is theoretically impossible to determine the two tree topologies from any amount of data under such circumstances. Here the question of identifiability is investigated for 2-tree mixtures of the 4-state group-based models, which are more relevant to DNA sequence data. Using algebraic techniques, we show that the tree parameters are identifiable for the JC and K2P models. We also prove that generic substitution parameters for the JC mixture models are identifiable, and for the K2P and K3P models obtain generic identifiability results for mixtures on the same tree. This indicates that the full phylogenetic signal remains in such mixtures, and that the 2-state symmetric result is thus a misleading guide to the behavior of other models.

I. INTRODUCTION

A basic question concerning any statistical model is whether a probability distribution arising from the model uniquely determines the parameters that produced it. If so, the parameters are said to be *identifiable*. Indeed, parameter identifiability is necessary for the consistency of inference.

In phylogenetics, it is especially important that the tree parameter of a model be identifiable, so that evolutionary histories can be consistently inferred. For basic models of character evolution along a tree, in which all sites behave independently and identically, identifiability of both the tree and continuous parameters is long-established. However, as phylogenetic models grow in complexity, it becomes increasingly difficult to analyze the models thoroughly enough to be certain this property is retained. Indeed, mixture models of all sorts present difficulties, though positive results have been obtained for models with a small number of classes evolving

on the same tree [5], and those with scaled Γ-distributed rates [1]. However, even for the GTR+Γ+I model, which is currently the most commonly used in DNA data analysis, it is yet to be proved that trees are identifiable.

Several recent works, including [25], [34], [24], and [23], considered 2-class mixture models in which the two classes evolve along possibly different topological trees. Such models could describe instances of horizontal transfer of genetic material between taxa, or incomplete lineage sorting in sequences composed of several concatenated genes. In particular, Matsen and Steel [24] showed that under the binary symmetric model of Cavender-Farris-Neyman, a 2-class mixture on a single 4-taxon tree can exactly 'mimic' a single class model on a different tree. Because of the small size of the state space in this model, its group-based structure, and the small size of the tree, explicit calculations were possible to fully analyze this situation. However, one should be cautious about extrapolating from this result to a pessimistic view about identifiability of similar phylogenetic mixtures. The mixture of [24] is an 11-parameter model producing a probability distribution in a 7-dimensional space, so it is certainly overparameterized. While this dimension count does not guarantee non-identifiability of the tree, it does explain why it might likely arise.

By either passing to models with larger state spaces, such as 4-state models appropriate to DNA, or by considering trees relating more taxa, the joint distribution of states at the leaves of the tree will be embedded in a larger dimensional space. Thus we might hope to avoid overparameterization issues through either of these modifications. As the analysis of real biological data typically involves both of these changes, these are the types of mixture models it is most desirable to understand.

Here we consider 2-class mixtures analogous to those in the works above, but for larger trees and/or state spaces.

We continue to work with group-based models, focusing primarily on those for DNA, so that we retain the powerful tool of the Fourier/Hadamard coordinate transformation.

We also make use of computational algebra software to perform calculations well beyond what could be done 'by hand.' Our results on identifiability are generally quite positive, and although these group-based models are still special cases, we believe they provide a better guide to the behavior of more realistic models than those of [24].

This paper is organized as follows. In Section II we introduce 2-tree mixture models and the identifiability problem in the algebraic setting. Background on group-based models is covered in Section III, from basic definitions through their

E.S. Allman and J.A. Rhodes are with the Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK, 99775-6660. E-mail: e.allman@alaska.edu, j.rhodes@alaska.edu

S. Petrović is with the Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL, 60607-7045. E-mail: petrovic@math.uic.edu

S. Sullivant is with the Department of Mathematics, North Carolina State University, Raleigh, NC, 27695. E-mail: smsulli2@ncsu.edu

presentation in terms of Fourier coordinates.

Section IV deals with identifiability of the tree parameters for Jukes-Cantor and Kimura 2-parameter mixture models on two trees. The main result, that tree parameters in such mixtures on at least 4 taxa are generically identifiable, is Theorem 10 and its corollary. Even with generic tree identifiability proved for 2-tree mixtures, a natural question is whether a single-class (unmixed) model can be distinguished from a 2-tree mixture. (This is not answered by the previous result, since while a single-class model is a special case of a 2-class model, it is non-generic.) We investigate this problem in Section V.

Finally, in Section VI, we turn to identifiability of the continuous parameters of these models, assuming the tree parameters are known. One feature of a part of our analysis is the use of computational algebra software to obtain some results *with very high probability*. Although technically these remain conjectures, lacking rigorous proof, the conclusions we draw from such calculations are highly reliable for theoretical reasons. While using calculations this way is familiar to applied algebraic geometers, this approach may be new to others, so we begin the section by explaining the reasoning informally. With this qualification, we establish the generic identifiability of continuous parameters for the JC model when either $n \geq 5$, or $n = 4$ and the trees are distinct. In the case of identical trees with $n \geq 5$ taxa, we give a fully rigorous argument for the three group-based models: JC, K2P, and K3P. An interesting non-identifiable case arises from the Jukes-Cantor mixtures on two identical 4-taxon trees.

Command files and instructions for verifying all our computations using the software Singular [16] can be found at the supplementary materials website for this paper [3]. We include both computations supporting our arguments, and those producing our examples.

We would, of course, prefer to push the work here beyond the group-based models, to include those more routinely used in current data analysis. It is possible, after all, that the group-based models are special enough that identifiability results for them do not carry over to more elaborate models. However, our current computational and theoretical tools are not sufficient for us to address questions for more general models.

II. PRELIMINARIES

Consider a phylogenetic model of k-state character change on n-taxon trees (e.g., for $k = 4$, the Jukes-Cantor model). We assume the taxa labelling the leaves are identified with $[n] = \{1, 2, , \ldots, n\}$. Then for each leaf-labelled tree T, there is a *parameterization map* ψ_T giving the joint distribution of states at the leaves of the tree T as functions of continuous parameters. With S_T denoting the continuous parameter space on T, which we assume is some full-dimensional subset of $\mathbb{R}^m$,

$$\psi_T : S_T \to \Delta^{k^n - 1},$$

where $\Delta^{k^n - 1} \subset [0, 1]^{k^n}$ is the probability simplex comprised of non-negative real vectors summing to 1.

Given such a model, the associated 2-tree mixture model has the following parameterization maps: For every pair of n-taxon trees T_1 and T_2 on the same taxa, let $S_{T_1, T_2} = S_{T_1} \times$

$S_{T_2} \times [0, 1]$ and

$$\psi_{T_1, T_2} : S_{T_1, T_2} \to \Delta^{k^n - 1},$$

be defined by

$$\psi_{T_1, T_2}(s_1, s_2, \pi) = \pi \, \psi_{T_1}(s_1) + (1 - \pi)\psi_{T_2}(s_2).$$

Here π is the *mixing parameter*, giving the proportion of i.i.d. sites that evolve along tree T_1.

We will only consider *algebraic models*, for which the maps ψ_T, and hence ψ_{T_1, T_2}, are defined by polynomial formulas. This is a small restriction, as many models (*e.g.*, standard continuous-time models) which are not polynomial can be embedded in ones that are (*e.g.*, the general Markov model). Algebraic models can be studied from the perspective of algebraic geometry [12], after extending ψ_T and ψ_{T_1, T_2} to complex polynomial maps, with images in $\mathbb{C}^{k^n}$. We refer to S_T and S_{T_1, T_2} as *stochastic parameter spaces*, to distinguish them from the *complex parameter spaces* of these extensions.

We denote by V_T the algebraic variety which is the Zariski closure of the image of ψ_T in the complex projective space $\mathbb{P}^{k^n - 1}$. (See [10], [17] for background in algebraic geometry.) Then the closure of the image of ψ_{T_1, T_2} is a variety called the *join* of V_{T_1} and V_{T_2}, denoted by

$$V_{T_1} * V_{T_2}.$$

The join can be described geometrically as the smallest variety containing all lines intersecting both V_{T_1} and V_{T_2}. In the case $T_1 = T_2$ the join is called the *secant* variety of V_{T_1}.

We use $\mathcal{M}_T$, and $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$, to denote the image of the parameterization maps when applied only to the stochastic parameter spaces S_T and S_{T_1, T_2}. Thus these denote the sets of all probability distributions arising from the parameterized models, and

$$\mathcal{M}_T \subsetneq V_T, \quad \mathcal{M}_{T_1} * \mathcal{M}_{T_2} \subsetneq V_{T_2} * V_{T_2}.$$

While $\mathcal{M}_T$ and $\mathcal{M}_{T_1} * \mathcal{M}_{T_2}$ are of course the objects of primary interest to phylogenetic applications, the larger complex varieties V_T and $V_{T_1} * V_{T_2}$ are more amenable to algebraic study.

Another parameterization of a dense subset of $V_{T_1} * V_{T_2}$, which we will also use, is

$$\phi_{T_1, T_2} : V_{T_1} \times V_{T_2} \times \mathbb{P}^1 \dashrightarrow V_{T_1} * V_{T_2},$$

which when restricted to an affine subset simply maps points on the two varieties to their convex sum using the third coordinate as a weight. (The dashed arrows indicates the map is only defined on a dense subset of the stated domain.) If $\pi \in \mathbb{C} \subset \mathbb{P}^1$, then

$$\psi_{T_1, T_2}(s_1, s_2, \pi) = \phi_{T_1, T_2}(\psi_{T_1}(s_1), \psi_{T_2}(s_2), \pi). \quad (1)$$

Associated to any algebraic variety V is the ideal $\mathcal{I} = \mathcal{I}(V)$ of polynomials that defines it; namely, a polynomial $f \in \mathcal{I}$ if, and only if, for any point $\mathbf{v} \in V$, $f(\mathbf{v}) = 0$. For a variety associated to a phylogenetic model, such polynomials give constraints that entries of a distribution of states at the leaves of a tree must satisfy if it arises from the given model. First introduced in phylogenetics by Cavender and Felsenstein [8]

and Lake [22], these polynomials are known as *phylogenetic invariants*, and have been studied extensively in many papers, including [13], [32], [33], [29], [20], [31], [6], [4].

For algebraic models, it is convenient to slightly weaken the notion of identifiability to *generic identifiability*. The word 'generic' is used to mean 'except on a proper algebraic subvariety' of the parameter space. Although it is sometimes possible to be explicit about this subvariety, we usually are not, since the key point in interpretation is that the subvariety is a closed set of Lebesgue measure 0 inside the larger set. Thus regardless of the precise subvariety involved, 'randomly' chosen points are generic with probability 1.

An additional issue for identifiability of 2-tree mixtures is class swapping: Interchanging the trees, along with their parameters, while replacing the mixing parameter π by $1 - \pi$, has no effect on the resulting distribution. Thus, a useful notion of identifiability must allow for this.

Definition 1: The tree parameters of the 2-tree mixture model are *generically identifiable* if, for any binary trees T_1, T_2 on the same set of taxa, and generic choices of s_1, s_2, π,

$$\psi_{T_1, T_2}(s_1, s_2, \pi) = \psi_{T_1', T_2'}(s_1', s_2', \pi')$$

implies $\{T_1, T_2\} = \{T_1', T_2'\}$.

Definition 2: The continuous parameters of a 2-tree mixture model on T_1 and T_2 are *generically identifiable* if for generic choices of s_1, s_2, π,

$$\psi_{T_1, T_2}(s_1, s_2, \pi) = \psi_{T_1, T_2}(s_1', s_2', \pi')$$

implies $(s_1, s_2, \pi) = (s_1', s_2', \pi')$, or, in the case where $T_1 = T_2$, $(s_1, s_2, \pi) = (s_2', s_1', 1 - \pi)$.

Let $K \subset [n]$ be a subset of the leaf set. For any tree T on n leaves, $T|_K$ will denote the induced subtree of T with leaf set K. Since marginalization onto leaf subsets is a linear map that preserves the mixture structure of a phylogenetic model, we obtain the following useful fact.

Lemma 3: Let T_1, T_2, T_3, T_4 be n-taxon trees, not necessarily distinct, and let $K \subseteq [n]$. If $V_{T_1|_K} * V_{T_2|_K} \nsubseteq V_{T_3|_K} * V_{T_4|_K}$, then $V_{T_1} * V_{T_2} \nsubseteq V_{T_3} * V_{T_4}$.

Proof: Marginalization to a fixed set K gives a linear map from $\mathbb{C}^{k^n}$ to $\mathbb{C}^{k^{|K|}}$, which sends V_T to $V_{T|_K}$ for any T. For any linear transformation A and any varieties, V, W, we have $A(V * W) = AV * AW$, because the mixture construction is a linear operation. Since for any sets S_1, S_2, and any map f, $f(S_1) \nsubseteq f(S_2)$ implies that $S_1 \nsubseteq S_2$, the lemma follows. ∎

III. GROUP-BASED PHYLOGENETIC MODELS

Group-based phylogenetic models will be the main subject of study in this paper, so we collect known results about these models, including their natural representation in Fourier coordinates.

Throughout we assume that all trees T are binary. We root T by picking an arbitrary edge, and introducing a root ρ as a distinguished node of degree 2 along it. Thus, every edge of T may be considered directed away from the root. To each node v in the tree, we associate a discrete random variable X_v with k states, and to each directed edge e in the tree,

we associate a Markov transition matrix M_e, describing the conditional probabilities of various state changes. We assume the reader is familiar with the usual assumptions of Markov processes on trees [27]. The *joint distribution of states at the leaves of T* may be computed (the image of ψ_T), once the root distribution and the collection of $\{M_e\}$ are specified. We refer to the entries of the root distribution and the Markov matrices as the *continuous parameters* of the model.

Definition 4: Let T be a binary tree rooted at ρ. Let G be a finite abelian group of order k, and identify its elements with the state space of the random variables X_v. Then a *group-based model* on T for the group G is a phylogenetic model with a uniform root distribution, and transition probabilities on each edge e satisfying

$$M_e(g, h) = f_e(g - h)$$

for some functions $f_e : G \to \mathbb{R}$.

Some standard examples of group-based models are the binary symmetric model, a.k.a. the Cavender-Farris-Neyman (CFN) model, which is associated to the group $\mathbb{Z}_2$; and the Jukes-Cantor (JC) model, the Kimura 2-parameter (K2P) model, and the Kimura 3-parameter (K3P) model, which are associated to the group $\mathbb{Z}_2 \times \mathbb{Z}_2$. With appropriate ordering of the state spaces, the transition matrices for these models have the following forms, respectively:

$$\begin{pmatrix} \alpha & \beta \\ \beta & \alpha \end{pmatrix}, \quad \begin{pmatrix} \alpha & \beta & \beta & \beta \\ \beta & \alpha & \beta & \beta \\ \beta & \beta & \alpha & \beta \\ \beta & \beta & \beta & \alpha \end{pmatrix},$$

$$\begin{pmatrix} \alpha & \beta & \gamma & \gamma \\ \beta & \alpha & \gamma & \gamma \\ \gamma & \gamma & \alpha & \beta \\ \gamma & \gamma & \beta & \alpha \end{pmatrix}, \quad \begin{pmatrix} \alpha & \beta & \gamma & \delta \\ \beta & \alpha & \delta & \gamma \\ \gamma & \delta & \alpha & \beta \\ \delta & \gamma & \beta & \alpha \end{pmatrix}.$$

Group-based models are subject to a remarkable linear change of coordinates, called the discrete Fourier, or Hadamard, transform [18], [19], [13], [32], [33]. After applying the Fourier transform the models are seen to be toric varieties [31]. In particular, the transformed image coordinates are given in terms of transformed domain coordinates by a monomial parameterization.

To make this parameterization explicit, henceforth let G be $\mathbb{Z}_2$ or $\mathbb{Z}_2 \times \mathbb{Z}_2$, and T an n-taxon tree. The Fourier coordinates for a group-based model are denoted $q_{g_1, \ldots, g_n}$, where $g_i \in G$ for all i. Let $\Sigma(T)$ be the set of splits induced by the edges of T. To each split $A|B \in \Sigma(T)$, we associate a set of parameters: $a_g^{A|B}$ where $g \in G$. The toric parameterization for the model is then given by:

$$q_{g_1, \ldots, g_n} = \begin{cases} \prod_{A|B \in \Sigma(T)} a_{\sum_{i \in A} g_i}^{A|B}, & \text{if } \sum_{i=1}^n g_i = 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$(2)$$

Note that by our choices of G, when $\sum_{i=1}^n g_i = 0$ we will have $\sum_{i \in A} g_i = \sum_{i \in B} g_i$ for any split $A|B$. Thus the formula above does not depend on the ordering of the sets in the splits.

To ease notation, we describe trees by omitting trivial splits associated to leaf edges (*i.e.*, those of the form $\{i\} \mid [n] \setminus \{i\}$).

When describing the toric parametrization of a group-based model, we abbreviate the parameters associated with the edge leading to leaf i by a_g^i.

For elements in the group $G = \mathbb{Z}_2 \times \mathbb{Z}_2$ associated to the Jukes-Cantor and Kimura models, we (arbitrarily) identify nucleotides with the group elements in the following way: $A = (0,0)$, $C = (0,1)$, $G = (1,0)$, and $T = (1,1)$. We illustrate these notions with an example on a 5-taxon tree.

Example 5: Let $T = \{12|345, 123|45\}$. The toric parameterization for the K3P model is given by formulas of the form:

$$q_{g_1 g_2 g_3 g_4 g_5} = a_{g_1}^1 a_{g_2}^2 a_{g_3}^3 a_{g_4}^4 a_{g_5}^5 a_{g_1+g_2}^{12|345} a_{g_1+g_2+g_3}^{123|45}$$

where $\sum_{i=1}^5 g_i = 0$, and $q_{g_1 g_2 g_3 g_4 g_5} = 0$ otherwise. For example,

$$q_{CCCTG} = a_C^1 a_C^2 a_C^3 a_T^4 a_G^5 a_A^{12|345} a_C^{123|45}.$$

For the JC and K2P models, the Fourier coordinates are described by simply imposing additional relationships on the continuous parameters.

Proposition 6: [13], [19] In Fourier coordinates, the K3P model on a tree T consists of all the Fourier vectors arising from representation (2) so that, for each split e, $a_A^e = 1$, and a_C^e, a_G^e, $a_T^e \in (0,1]$. The K2P model is the submodel of the K3P model satisfying additionally, that for all splits e, $a_G^e = a_T^e$. The JC model is the submodel of the K2P model satisfying additionally, that for all splits e, $a_C^e = a_G^e = a_T^e$.

Significantly for the work later in this article, the Fourier transform is a *linear* change of coordinates. Thus the operation of taking tree mixtures commutes with the Fourier transform, which allows us to naturally represent the mixture models we consider in Fourier coordinates. Though these mixture models are not toric, we still gain insight from this viewpoint.

To close this section, we make several comments regarding some combinatorial aspects of Fourier coordinates for group-based models. As linear invariants are crucial to the arguments below, we first discuss enumeration of distinct Fourier coordinates, and computations of the dimension of the space of linear invariants for a model. In closing, we illustrate some useful combinatorial mnemonics for working with Fourier coordinates and identifying invariants. Although these devices are likely familiar to experts in group-based models, we hope our exposition will be useful to those less familiar with these models.

The zero set of the linear invariants for any variety $V \subset \mathbb{P}^n$ is the smallest linear subspace of $\mathbb{P}^n$ containing V. This set is called the *span of* V, $\mathrm{Span}(V)$. The span of a finite collection of varieties is defined similarly, as the span of their union. Note that, by the join construction, it is immediate that $\mathrm{Span}(V * W) = \mathrm{Span}(V, W)$.

For group-based models on an n-taxon tree (that is, an *unmixed* model), the number of distinct Fourier coordinates is precisely the dimension of the span of the model, as there are no linear relations between distinct monomials. This establishes:

Proposition 7: For the CFN and K3P models, there are no non-trivial linear invariants. The number of distinct Fourier coordinates is 2^{n-1} for the CFN model and 4^{n-1} for the K3P model.

As our method of investigation of mixture models in Section IV depends upon the existence of linear invariants, we thus focus on the JC and K2P group-based models.

Steel and Fu [29] computed the dimension of $\mathrm{Span}(V_T)$ for the JC model. Hendy and Penny [20] performed a similar computation for the K2P model.

Theorem 8: [29], [20] Let T be an n-taxon binary tree. Then, for the JC model on T, the number of distinct Fourier coordinates is the Fibonacci number F_{2n-2}, satisfying the recurrence

$$F_0 = 1, \; F_1 = 1, \; F_i = F_{i-1} + F_{i-2}.$$

That is, for the JC model $\mathrm{Span}(V_T)$ has dimension F_{2n-2}.

For the K2P model on T, the number of distinct Fourier coordinates is H_n, satisfying the recurrence:

$$H_1 = 1, \; H_2 = 3, \; H_i = 4H_{i-1} - 2H_{i-2}.$$

That is, for the K2P model $\mathrm{Span}(V_T)$ has dimension H_n.

Fourier coordinates for group-based models have useful combinatorial representatives in terms of labelled or colored versions of the underlying tree T. For this representation, we associate a color to each of the different parameter classes in the model. For example, in the JC model, there are two parameter classes: the A class (grey), and the $\{C, G, T\}$ class (black). With this choice of colors, in the parametric description of $q_{g_1,\ldots,g_n}$ if a parameter $a_A^{B|B'}$ occurs, then we color the edge corresponding to the split $B|B'$ grey. If a parameter $a_C^{B|B'}$, $a_G^{B|B'}$, or $a_T^{B|B'}$ occurs, we color the edge corresponding to the split $B|B'$ black. As shown in [29], this establishes a correspondence between distinct Fourier coefficients for the JC model and subforests of T with the same leaf set $[n]$.

The color-coding of the underlying tree works similarly for the K2P model; here we have three classes of parameters, the A-class, the C-class, and the $\{G, T\}$-class, and hence use three colors.

Example 9: Continuing Example 5, the colored diagram corresponding to the Fourier coordinate q_{CCCTG} for the JC model is shown in Figure 1.
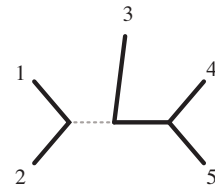

Fig. 1. JC-coloring for q_{CCCTG}. (Key: A-class = dashed grey, $\{C, G, T\}$-class = black)

For the K2P model, the same Fourier coordinate is represented by the tri-colored tree of Figure 2.

These diagrams are useful for determining the invariants that a particular group-based model satisfies. For instance, one phylogenetic invariant for the JC model on the tree T with split $12|34$ is given in Fourier coordinates by

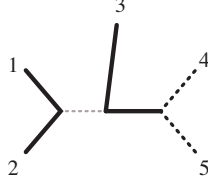$$q_{CTGA} q_{ACTG} = q_{CGCG} q_{ACCA}. \qquad (3)$$

Fig. 2. K2P-coloring for q_{CCCTG}. (Key: A-class = dashed gray, C-class = black, $\{G, T\}$-class = dashed black)

This relationship may be represented in pictorial form by the diagram in Figure 3.
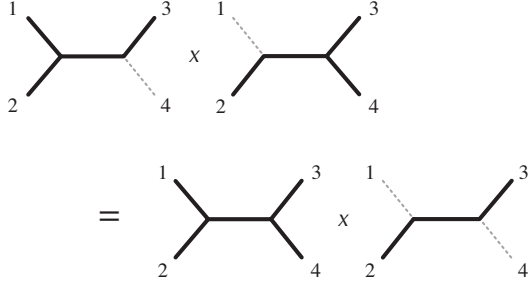

Fig. 3. Pictorial view of invariant (3) for the JC model on T.

IV. IDENTIFIABILITY OF TREE PARAMETERS

The goal of this section is to prove that the tree parameters are generically identifiable for 2-tree JC and K2P mixture models on at least 4 taxa. For the complex varieties associated to the models this is formalized as follows:

Theorem 10: Suppose T_1, T_2, T_3, T_4 are binary trees, not necessarily distinct, on $n \geq 4$ taxa, and consider the 2-tree mixture varieties for the JC and K2P models. If $\{T_1, T_2\} \neq \{T_3, T_4\}$, then $V_{T_1} * V_{T_2} \not\subseteq V_{T_3} * V_{T_4}$.

If $\{T_1, T_2\} \neq \{T_3, T_4\}$, then the noncontainment of the irreducible varieties $V_{T_1} * V_{T_2}$ and $V_{T_3} * V_{T_4}$ in one another shows their intersection is a proper subvariety of strictly lower dimension. The preimage of this intersection under the complex parameterization map is thus a proper subvariety of the parameter space. Since the subset of stochastic parameters is Zariski dense in the complex parameter space, those stochastic parameters mapping to the intersection also lie in a closed set of Lebesgue measure 0. Thus we obtain the main result of the section:

Corollary 11: For the 2-tree JC and K2P mixture models on at least 4 taxa the tree parameters are generically identifiable for either stochastic or complex parameters.

The proof of Theorem 10 proceeds in three parts. First, we show that the two tree parameters when $T_1 = T_2$ are identifiable. Then we focus on the quartet trees, constructing a linear invariant that completes the proof of the result for quartet tree mixtures. Finally, we combine our quartet results, the six-to-infinity theorem of Matsen, Mossel, and Steel [23], and a linear invariant for 6-taxon tree mixtures, to deduce identifiability of trees for 2-tree mixtures on an arbitrary number of leaves.

A. 2-Tree Mixture with $T_1 = T_2$

In this section, we focus on a mixture of a tree with itself; that is, we study the secant variety $V_T * V_T$. We show that $V_T * V_T$ can be distinguished from any 2-tree mixture variety $V_{T_1} * V_{T_2}$, provided T_1 and T_2 are not both T.

Proposition 12: Let T_1, T_3, T_4 be three binary trees, not necessarily distinct, with $n \geq 4$ leaves, such that $\{T_1\} \neq \{T_3, T_4\}$. Then under both the JC and K2P 2-tree mixture models, $V_{T_1} * V_{T_1} \not\subseteq V_{T_3} * V_{T_4}$ and $V_{T_3} * V_{T_4} \not\subseteq V_{T_1} * V_{T_1}$.

Proof: Assume $T_3 \neq T_1$. By [29] and [20], for both the unmixed JC and K2P models, there exists a linear invariant $l \in \mathcal{I}(V_{T_1}) \setminus \mathcal{I}(V_{T_3})$. Since the set of linear invariants of V_{T_1} and $V_{T_1} * V_{T_1}$ coincide (the varieties have the same span), there exists a linear invariant $l \in \mathcal{I}(V_{T_1} * V_{T_1}) \setminus \mathcal{I}(V_{T_3})$. Hence, $V_{T_3} \not\subseteq V_{T_1} * V_{T_1}$. Now since $V_{T_3} \subseteq V_{T_3} * V_{T_4}$, it follows that $V_{T_3} * V_{T_4} \not\subseteq V_{T_1} * V_{T_1}$.

It remains to show $V_{T_1} * V_{T_1} \not\subseteq V_{T_3} * V_{T_4}$. In fact, it is enough to show that

$$\dim V_{T_3} * V_{T_4} \leq \dim V_{T_1} * V_{T_1}. \tag{4}$$

Indeed, if this inequality holds strictly, the claim is obvious. If, on the other hand, the dimensions are equal, then since both of the joins are irreducible varieties, $V_{T_1} * V_{T_1} \subseteq V_{T_3} * V_{T_4}$ would imply equality of varieties, contradicting the anti-containment already established.

Now a simple bound on the dimension of a join, coming from its natural parameterization, is

$$\dim V * W \leq \dim V + \dim W + 1,$$

where the quantity on the right is called the *expected dimension* when it is no larger than the dimension of the ambient space. In the case of the JC model, $\dim(V_{T_3}) = \dim(V_{T_4}) = 2n - 3$, which shows $\dim(V_{T_3} * V_{T_4}) \leq 4n - 5$. Similarly, $\dim(V_{T_3} * V_{T_4}) \leq 8n - 11$ for the K2P model.

To complete the proof of the claim of Proposition 12 for the JC model by establishing inequality (4), it suffices to show the secant variety has the expected dimension, as given by the following:

Lemma 13: If T is an n-taxon binary tree then $\dim V_T * V_T = 4n - 5$ for the JC model.

As the proof of this lemma is more involved, we defer it until after our current argument. For the K2P model, we prove below a weaker claim.

Lemma 14: If T is a 4-taxon binary tree, then $\dim V_T * V_T = 21$ for the K2P model.

This is sufficient to complete the proof of the claim of Proposition 12 in the K2P case for 4-taxon trees. Larger trees are then treated by considering marginalizations to induced 4-taxon trees: Choose a set K of 4 taxa for which the induced quartet trees $T_1|_K, T_3|_K, T_4|_K$ are not all the same, and then apply Lemma 3. ∎

To prove Lemma 13, we make use of a special case of the tropical secant varieties theory of Draisma [11] and the fact that the varieties V_T are toric varieties. To explain Draisma's result (Theorem 15 below), we introduce some background material on toric varieties and convex geometry.

Recall that a toric variety is specified as the image of a polynomial map, each of whose coordinate functions is a monomial. As a monomial is of the form $x^u = x_1^{u_1} x_2^{u_2} \cdots x_d^{u_d}$, we associate to each monomial a non-negative integer vector u. To a toric variety, we associate a collection of non-negative integer vectors $A \subset \mathbb{N}^d$, one vector for each monomial appearing in the parameterization. We also identify A with a matrix whose columns are the given set. The toric variety is often denoted V_A. Algebraic and geometrical properties of toric varieties are reflected in corresponding properties of the vector configuration A [14], [30].

By a hyperplane in $\mathbb{R}^d$, we mean a linear hypersurface $H = \{x \in \mathbb{R}^d : c^T x = e\}$. The complement $\mathbb{R}^d \setminus H$ consists of two connected components, which we denote by H^+ and H^-.

Theorem 15 ([11]): Let V_A be a projective toric variety, with corresponding set of exponent vectors $A \subseteq \mathbb{N}^d$. Suppose that A has rank r, so that $\dim V_A = r - 1$. Let H be a hyperplane not intersecting A. Let $A^+ = A \cap H^+$ and $A^- = A \cap H^-$. Then $\dim V_A * V_A \geq \operatorname{rank} A^+ + \operatorname{rank} A^- - 1$. In particular, if there exists an H such that $\operatorname{rank} A^+ = \operatorname{rank} A^- = \operatorname{rank} A$, then V_A has the expected dimension.

Proof: [Proof of Lemma 13] To apply Theorem 15, we must investigate the vector configurations A associated with the Jukes-Cantor model and find a hyperplane with the desired properties. For the Jukes-Cantor model on a binary tree, each distinct Fourier coordinate corresponds to a subforest F of the tree, and the corresponding monomial has the form

$$\prod_{e \in F} a_C^e \prod_{e \notin F} a_A^e.$$

(Here we consider the a_A^e as variables, rather than setting them to be 1. This simply homogenizes our parameterization.) The vector corresponding to F is in $\mathbb{N}^{4n-6}$; specifically, $u^F = (x_e, y_e)_{e \in \Sigma(T)}$ such that $x_e = 1$ and $y_e = 0$ if $e \in F$, and $x_e = 0$ and $y_e = 1$ if $e \notin F$. For example, in the case that $n = 4$, and T is the tree with nontrivial split $12|34$, then, after removing repeated columns, A consists of the columns of the 10×13 matrix:

$$\begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

The first 2 rows here correspond to the x_e for edges in one cherry on the tree, the next 2 to x_e for edges in the other cherry, and the 5th to x_e for the central edge; the last 5 correspond to the y_e, with edges in the same order. Thus, the first column corresponds to the empty forest, the second to the first cherry, and so on.

Consider the hyperplane $H = \{(x_e, y_e) \in \mathbb{R}^{4n-6} : \sum_{e \in \Sigma(T)} x_e = |\Sigma(T)| - 3/2\}$. This means that the vectors on one side of the partition will correspond to subforests of T having at most one edge of T missing. The subforests on

the other side will have at least two edges missing. Call the first set of vectors A^+ and the second set of vectors A^-. In the matrix above, A^+ consists of the last six columns and A^- consists of the first seven columns.

The first set A^+ contains exactly $|\Sigma(T)| + 1$ vectors, since the tree itself is a subforest and removing any edge always produces a subforest. This set thus forms the vertices of a simplex of dimension equal to the number of edges, so A^+ has rank $2n - 2$.

The second set, A^-, contains the empty graph and all paths between pairs of vertices. If we restrict attention to only those vectors corresponding to the paths between pairs of vertices, this gives us the exponent vectors of the toric varieties corresponding to toric degenerations of the Grassmannian [28], which has rank $2n - 3$. Adding the vector corresponding to the empty subforest increases the rank by one. ∎

Proof: [Proof of Lemma 14] To apply Theorem 15, we must investigate the vector configurations A associated with the K2P model and find a hyperplane with the desired properties. For $n = 4$, there are $H_4 = 34$ distinct Fourier coordinates. We focus on the tree with the unique nontrivial split $12|34$. Each monomial in the Fourier parameterization has the form

$$a_{g_1}^1 a_{g_2}^2 a_{g_3}^3 a_{g_4}^4 a_{g_1+g_2}^{12|34},$$

where $a_G^e = a_T^e$ for all splits e. This implies that the matrix A is a 15×34 matrix. The coordinates on $\mathbb{R}^{15}$ are x_e, y_e, z_e, where for a given edge e,

$$(x_e, y_e, z_e) = \begin{cases} (1, 0, 0) & \text{if } g_e = A \\ (0, 1, 0) & \text{if } g_e = C \\ (0, 0, 1) & \text{if } g_e = G, T. \end{cases}$$

Since the model has dimension 10, we see that the matrix A has rank 11.

Now consider the hyperplane $H = \{(x_e, y_e, z_e) \in \mathbb{R}^{15} : \sum_{e \in \Sigma(T)} y_e + z_e = 7/2\}$. A direct calculation shows that this partitions A into A^+ and A^- each with rank 11, and completes the proof. ∎

B. Linear Invariants for Quartet Mixtures

We next focus on quartet trees. The three fully-resolved quartet trees will be indicated by their non-trivial splits: $T_{12|34}$, $T_{13|24}$, and $T_{14|23}$. The main result of this section is that linear invariants can generically identify 2-quartet mixtures.

Lemma 16: For both the JC and K2P models, the linear polynomial

$$f = q_{GGGG} + q_{GTGT} - q_{GGTT} - q_{GTTG}$$

satisfies $f \in \mathcal{I}(V_{T_{12|34}} * V_{T_{14|23}}) \setminus \mathcal{I}(V_{T_{13|24}})$.
Note the lemma further implies

$$f \in \mathcal{I}(V_{T_{12|34}} * V_{T_{14|23}}) \setminus \mathcal{I}(V_{T_{13|24}} * V_{T_{14|23}}).$$

Combining this with Proposition 12, we deduce a first case of Theorem 10:

Corollary 17: The case $n = 4$ of Theorem 10 holds.

Proof: [Proof of Lemma 16] Denote the parameters for tree $T_{12|34}$ by a and the parameters for $T_{14|23}$ by b. We must

show that $f = 0$ whenever we substitute for the q's the parameterization for the mixture model. One checks that:

$$q_{GGGG} = \pi a_G^1 a_G^2 a_G^3 a_G^4 a_A^{12|34} + (1-\pi) b_G^1 b_G^2 b_G^3 b_G^4 b_A^{14|23},$$

$$q_{GTGT} = \pi a_G^1 a_T^2 a_G^3 a_T^4 a_C^{12|34} + (1-\pi) b_G^1 b_T^2 b_G^3 b_T^4 b_C^{14|23},$$

$$q_{GGTT} = \pi a_G^1 a_G^2 a_T^3 a_T^4 a_A^{12|34} + (1-\pi) b_G^1 b_G^2 b_T^3 b_T^4 b_C^{14|23},$$

$$q_{GTTG} = \pi a_G^1 a_T^2 a_T^3 a_G^4 a_C^{12|34} + (1-\pi) b_G^1 b_T^2 b_T^3 b_G^4 b_A^{14|23}.$$

Since for the K2P and JC models $a_G^e = a_T^e$, $b_G^e = b_T^e$ for all e, these formulae show $f = 0$, as can be checked using color-codes trees such as in Section III. Thus $f \in \mathcal{I}(V_{T_{12|34}} * V_{T_{14|23}})$.

On the other hand, for the tree $T_{13|24}$ we have:

$$q_{GGGG} = c_G^1 c_G^2 c_G^3 c_G^4 c_A^{13|24}$$

$$q_{GTGT} = c_G^1 c_T^2 c_G^3 c_T^4 c_A^{13|24}$$

$$q_{GGTT} = c_G^1 c_G^2 c_T^3 c_T^4 c_C^{13|24}$$

$$q_{GTTG} = c_G^1 c_T^2 c_T^3 c_G^4 c_C^{13|24}$$

Even though in the JC model $c_C^e = c_G^e = c_T^e$ and $c_A^e = 1$ for all e, f is not identically zero when evaluated at these expressions. Thus $f \notin \mathcal{I}(V_{T_{13|24}})$ for the JC model, and hence also for the K2P model. ∎

C. From Quartets to Sextets and Beyond

Identifiability of quartet mixtures can be used to show identifiability for larger trees by marginalization of tree models and their mixtures. However, it is not, in general, possible to identify two trees from the union of their sets of induced quartet trees. Thus this approach requires some care. That all difficulties arise from trees of at most 6 taxa is the content of the following combinatorial theorem of Matsen, Mossel, and Steel.

Theorem 18 (Six-to-Infinity Theorem): [23] Suppose that the tree parameters T_1, T_2 are identifiable for a 2-tree phylogenetic mixture model for binary trees with six leaves. Then tree parameters are identifiable for binary trees with ≥ 6 leaves.

Combining the results of Corollary 17 and Lemma 3, we have that $V_{T_1} * V_{T_2} \not\subseteq V_{T_3} * V_{T_4}$, if there is a four element subset $Q \subseteq [n]$ such that $\{T_1|_Q, T_2|_Q\} \neq \{T_3|_Q, T_4|_Q\}$.

It remains to show that $V_{T_1} * V_{T_2} \not\subseteq V_{T_3} * V_{T_4}$ for pairs of trees such that $\{T_1|_Q, T_2|_Q\} = \{T_3|_Q, T_4|_Q\}$ for *all* four element subsets $Q \subseteq [n]$. Let $\mathcal{Q}(T_i, T_j)$ denote the multiset of all quartet trees $T_i|_Q, T_j|_Q$ induced by T_i and T_j. We say two pairs of trees T_1, T_2 and T_3, T_4 are *quartet-matched* if $\mathcal{Q}(T_1, T_2) = \mathcal{Q}(T_3, T_4)$.

Proposition 19: For $n = 5$ leaves, any two quartet-matched pairs of trees T_1, T_2 and T_3, T_4 has $\{T_1, T_2\} = \{T_3, T_4\}$. For $n = 6$ leaves, every quartet-matched pair of trees with $\{T_1, T_2\} \neq \{T_3, T_4\}$ is equivalent, up to $\mathfrak{S}_6$ symmetry, to the pairs defined by

$$T_1 = \{12|3456, 123|456, 1234|56\},$$
$$T_2 = \{13|2456, 123|456, 1235|46\}, \quad (5)$$
$$T_3 = \{13|2456, 123|456, 1234|56\},$$
$$T_4 = \{12|3456, 123|456, 1235|46\}.$$

Proof: Fix two binary trees T_1, T_2 with n leaves. If the trees are identical, the result is clear, so we assume throughout $T_1 \neq T_2$.

Consider first $n = 5$.

If leaves j, k form a cherry in T_i, then they will also form a cherry in all 3 quartet trees including j, k induced from T_i. On the other hand, if they do not form a cherry in T_i, then they will form a cherry in either 0 or 1 of these 3 induced quartet trees. Thus by counting the elements of the multiset $\mathcal{Q}(T_1, T_2)$ with each possible cherry j, k, we can determine which cherries occur in both trees (count 6), which occur in exactly one tree (count 3 or 4), and which occur in no trees (count 0, 1, or 2).

If a cherry occurs in both trees, suppose it is $\{1, 2\}$. Then from considering the quartets on $\{2, 3, 4, 5\}$ both T_1 and T_2 are determined.

If the two trees have no cherry in common, then we know the 4 distinct cherries that occur in the 2 trees. If only 4 taxa occur in these 4 cherries, then we may uniquely pair them according to their compatibility, and the two 5-taxon trees T_1 and T_2 are determined. If all 5 taxa occur in these cherries, since the cherries are distinct we may assume they are $\{1, 2\}$ and $\{3, 4\}$ (from one tree), and $\{1, 5\}$ and $\{2, 3\}$ (from the other), though we initially do not know which come from which tree. However, we again see that these can be uniquely paired for compatibility, and thus T_1 and T_2 are determined.

Now consider $n = 6$.

By the $n = 5$ case, we may determine the multiset $\mathcal{F} = \mathcal{F}(T_1, T_2)$ of all 5-taxon induced trees from T_1 and T_2, so we work with it instead of $\mathcal{Q} = \mathcal{Q}(T_1, T_2)$.

By counting cherries in $\mathcal{F}$, we may determine those possible cherries that occur in both trees (count 8), exactly one tree (count 4 or 5), or no trees (count 0,1, or 2). If a cherry occurs in both trees, suppose it is $\{1, 2\}$. Then from considering the 5-taxon trees on $\{2, 3, 4, 5, 6\}$ both T_1 and T_2 are determined.

For the reminder of the proof, we assume the trees have no cherry in common. Thus either 4, 5, or 6 distinct cherries occur in T_1 and T_2. In the case of 6 distinct cherries, compatibility of cherries determines T_1 and T_2.

In the case of 5 distinct cherries, one of the trees must be symmetric, and the other a caterpillar. Either compatibility of cherries determines the symmetric tree (in which case both trees are determined by removing the quartets from this tree from $\mathcal{Q}$ and using the remaining ones to construct the second tree), or we may assume the 5 cherries have the form $\{1, 2\}, \{3, 4\}, \{5, 6\}$ (from one tree) and $\{1, 3\}, \{2, 4\}$ (from the other), though of course we do not know which come from which tree. Since the cherry $\{5, 6\}$ is identified by this, consider the two elements of $\mathcal{F}$ on $\{1, 2, 4, 5, 6\}$. As $\{5, 6\}$ is a cherry in only one of these trees, and that one also has $\{1, 2\}$ as its other cherry, this identifies $\{1, 2\}$. Thus $\{3, 4\}$ is also known, and thus one, and hence both, of the trees are determined.

In the case of 4 distinct cherries, both T_1 and T_2 are caterpillars. We first investigate whether we can determine which pairs of cherries occur on the same tree T_i. Since at least 2 of the 4 cherries must be incompatible, let these be

$\{1,2\}$ and $\{1,3\}$. Either compatibility determines which other cherries these are paired with, or the remaining cherries have the form $\{i,j\}$ with $i,j \in \{4,5,6\}$, and we may assume the cherries are $\{4,6\}$ and $\{5,6\}$.

If compatibility determined the cherries on T_1 as $\{1,2\}$ and $\{i,j\}$, and those on T_2 as $\{1,3\}$ and $\{k,l\}$, then we may assume $j \neq 3$. Then the two elements of $\mathcal{F}$ on all taxa but j can be matched with the T_i depending on whether they display the cherry $\{1,2\}$ or $\{1,3\}$. This determines T_1, and hence T_2 as well.

This leaves only the case where the 4 cherries are $\{1,2\}$, $\{1,3\}$, $\{4,6\}$ and $\{5,6\}$, which may be paired two ways. Considering the two elements of $\mathcal{F}$ on $\{1,2,3,4,5\}$, exactly one must contain the cherry $\{1,2\}$. If the other cherry in this 5-taxon tree is $\{3,5\}$, then this determines T_1 as $\{12|3456, 124|356, 1234|56\}$, and hence T_2 is determined as well. Similarly, if the second cherry in the 5-taxon tree is $\{3,4\}$, then T_1 and T_2 are again uniquely determined. If the second cherry is $\{4,5\}$, however, T_1 may be either $\{12|3456, 123|456, 1234|56\}$ or $\{12|3456, 123|456, 1235|46\}$. Considering the element of $\mathcal{F}$ on $\{1,2,3,4,5\}$ that contains cherry $\{1,3\}$, we likewise obtain two unique trees except in the case where the second cherry is $\{4,5\}$, in which case T_2 could be either $\{13|2456, 123|456, 1234|56\}$ or $\{13|2456, 123|456, 1235|46\}$. Finally, since only one of T_1 and T_2 can have cherry $\{5,6\}$, the only remaining ambiguous case is that described in the statement of the Proposition. ■

Lemma 20: Consider the trees T_1, T_2, T_3, and T_4 in equations (5) from Proposition 19. Define the linear polynomial

$$f = q_{GGGGGG} + q_{GTTTTG} - q_{GTGGTG} - q_{GGTTGG}.$$

Then, for the JC and K2P models, f satisfies

$$f \in \mathcal{I}(V_{T_1} * V_{T_2}) \setminus \mathcal{I}(V_{T_i}), \quad i \in \{3,4\}.$$

In particular, $V_{T_i} \not\subseteq V_{T_1} * V_{T_2}$, $i \in \{3,4\}$.

Proof: By symmetry of the relationship between trees T_3 and T_4 to that of T_1 and T_2, it suffices to prove the statement in the case that $i = 3$. First, we will show that $f \in \mathcal{I}(V_{T_1} * V_{T_2})$. Denote by a's and b's the parameters of the trees T_1 and T_2, respectively. One checks that

$$q_{GGGGGG} = \pi a_G^1 a_G^2 a_G^3 a_G^4 a_G^5 a_G^6 a_A^{12|3456} a_G^{123|456} a_A^{1234|56}$$
$$+ (1-\pi) b_G^1 b_G^2 b_G^3 b_G^4 b_G^5 b_G^6 b_A^{13|2456} b_G^{123|456} b_A^{1235|46},$$

$$q_{GTTTTG} = \pi a_G^1 a_T^2 a_T^3 a_T^4 a_T^5 a_G^6 a_C^{12|3456} a_G^{123|456} a_C^{1234|56}$$
$$+ (1-\pi) b_G^1 b_T^2 b_T^3 b_T^4 b_T^5 b_G^6 b_C^{13|2456} b_G^{123|456} b_C^{1235|46},$$

$$q_{GTGGTG} = \pi a_G^1 a_T^2 a_G^3 a_G^4 a_T^5 a_G^6 a_C^{12|3456} a_T^{123|456} a_C^{1234|56}$$
$$+ (1-\pi) b_G^1 b_T^2 b_G^3 b_G^4 b_T^5 b_G^6 b_A^{13|2456} b_T^{123|456} b_A^{1235|46},$$

$$q_{GGTTGG} = \pi a_G^1 a_G^2 a_T^3 a_T^4 a_G^5 a_G^6 a_A^{12|3456} a_T^{123|456} a_A^{1234|56}$$
$$+ (1-\pi) b_G^1 b_G^2 b_T^3 b_T^4 b_G^5 b_G^6 b_C^{13|2456} b_T^{123|456} b_C^{1235|46}.$$

Recall that for the JC and K2P models, we have $a_G^e = a_T^e$ and $b_G^e = b_T^e$ for all e. Therefore, $f \in \mathcal{I}(V_{T_1} * V_{T_2})$.

On the other hand, if we denote the parameters for the tree T_3 by c's, we have that:

$$q_{GGGGGG} = c_G^1 c_G^2 c_G^3 c_G^4 c_G^5 c_G^6 c_A^{13|2456} c_G^{123|456} c_A^{1234|56},$$

$$q_{GTTTTG} = c_G^1 c_T^2 c_T^3 c_T^4 c_T^5 c_G^6 c_C^{13|2456} c_G^{123|456} c_C^{1234|56},$$

$$q_{GTGGTG} = c_G^1 c_T^2 c_G^3 c_G^4 c_T^5 c_G^6 c_A^{13|2456} c_T^{123|456} c_C^{1234|56},$$

$$q_{GGTTGG} = c_G^1 c_G^2 c_T^3 c_T^4 c_G^5 c_G^6 c_C^{13|2456} c_T^{123|456} c_A^{1234|56}.$$

Although for the JC model $c_C^e = c_G^e = c_T^e$ and $c_A^e = 1$ for all e, the linear polynomial f evaluated at the expressions above does not give the zero polynomial. Therefore $f \notin \mathcal{I}(V_{T_3})$ for the JC, and hence for the K2P, model. ■

Finally, we pull together all of the results in this section in the proof of the main Theorem on tree identifiability:

Proof: [Proof of Theorem 10] If the trees relate only 4 taxa, Corollary 17 provides the claim. By Theorem 18, it is now enough to consider cases with $n = 5, 6$.

If $\{T_1, T_2\}$ and $\{T_3, T_4\}$ are not quartet-matched, then there is a quartet Q of taxa such that $\{T_1|_Q, T_2|_Q\} \neq \{T_3|_Q, T_4|_Q\}$. Thus by Corollary 17 and Lemma 3, the claim follows.

If $\{T_1, T_2\}$ and $\{T_3, T_4\}$ are quartet-matched, then by Proposition 19, up to symmetry, we need only consider the case described by equations (5). But then Lemma 20 implies

$$\mathcal{I}(V_{T_3} * V_{T_4}) \setminus \mathcal{I}(V_{T_1} * V_{T_2})$$

contains a linear invariant, so

$$V_{T_1} * V_{T_2} \not\subseteq V_{T_3} * V_{T_4}.$$

■

V. COMPARING 2-TREE MIXTURES WITH UNMIXED MODELS

In this section, we report on preliminary investigations on distinguishing unmixed models from 2-tree mixtures. More precisely, we study the following question: For which triples of trees T_1, T_2, T_3 is $V_{T_3} \not\subseteq V_{T_1} * V_{T_2}$? We have already used instances of this in establishing generic identifiability of trees in the 2-tree mixture model, but our earlier work does not yield a general answer.

That we can distinguish a single-class, unmixed model V_{T_3} from a 2-tree mixture model $V_{T_1} * V_{T_2}$, as long as T_3 is not too closely related to T_1 and T_2 is easily shown, however. Indeed, Lemma 16 and a variant of Lemma 3 imply:

Proposition 21: If there is a four-element set $Q \subseteq [n]$ such that $T_3|_Q \notin \{T_1|_Q, T_2|_Q\}$ then $V_{T_3} \not\subseteq V_{T_1} * V_{T_2}$.

The smallest instance of a tree T_3 all of whose quartet trees arise from T_1 and T_2 occurs with $n = 5$ leaves for the triple:

$$\begin{aligned} T_1 &= \{12|345, 123|45\}, \\ T_2 &= \{13|245, 134|25\}, \\ T_3 &= \{123|45, 13|245\}. \end{aligned} \qquad (6)$$

In fact, this example is unique up to the action of $\mathfrak{S}_5$ on leaf labels.

We performed a computation using the computer algebra program Singular [16], which rigorously verified the following:

Theorem 22: For the three 5-taxon trees T_1, T_2, T_3 in (6), under the JC model,

$$V_{T_3} \subseteq V_{T_1} * V_{T_2}.$$

Proof: We explain the approach behind our computation.

All of the JC varieties V_T for an n-leaf tree are invariant under an action of the torus $(\mathbb{C}^*)^n$. This action arises from rescaling the pendant edge parameters. That is, if $q \in V_T$ and $\lambda \in (\mathbb{C}^*)^n$, then for any subforest F of T the Fourier coordinate q_F is transformed as $(\lambda \cdot q)_F = q_F \prod_{e \in L(F)} \lambda_e$, where $L(F)$ is the set of pendant edges appearing in F. Since $\lambda \cdot q \in V_T$, it suffices to prove the claimed containment in the theorem in the case where all pendant edge parameters on the tree T_3 are set to 1.

Let V and W be two varieties. Note that $V \subseteq W$, is equivalent to $\mathcal{I}(W) \subseteq \mathcal{I}(V)$. This containment of ideals holds if, and only if, $\mathcal{I}(W) + \mathcal{I}(V) = \mathcal{I}(V)$, which we use to speed up computations. Hence, it suffices to show that

$$\mathcal{I}(V_{T_1} * V_{T_2}) + \mathcal{I}(V) = \mathcal{I}(V) \qquad (7)$$

where V is the subvariety of V_{T_3} where all the pendant edge parameters are set to 1.

Finally, though in principle it is possible to compute $\mathcal{I}(V_{T_1} * V_{T_2})$ directly, it is beyond current capabilities. However, an alternative approach to join ideals uses elimination: if $I, J \subseteq \mathbb{C}[q]$ are two ideals, their join ideal is

$$I * J = (I(q') + J(q - q')) \cap \mathbb{C}[q]$$

where $I(q')$ is the ideal I with variables q_i' substituted for variables q_i, and $J(q - q')$ is ideal J with $q_i - q_i'$ substituted for q_i. Hence, we can test (7) by testing if

$$\mathcal{I}(V) = (\mathcal{I}(V_{T_1})(q') + \mathcal{I}(V_{T_2})(q - q') + \mathcal{I}(V)(q)) \cap \mathbb{C}[q].$$

This statement is verified by the code we provide in the supplementary materials [3]. ∎

Theorem 22 raises as many questions as it answers. First, note that it is a statement about complex varieties, and leaves open the possibility that $\mathcal{M}_{T_3} \not\subseteq \mathcal{M}_{T_1} * \mathcal{M}_{T_2}$. We investigated this computationally as follows, using code available in [3]. Choosing random JC parameters on T_3, we repeatedly produced a point in $V_{T_3} \subseteq V_{T_1} * V_{T_2}$, thus obtaining a sample with high probability of exhibiting generic behavior. For each such point, we then produced a system of algebraic equations whose solutions would give mixture parameters on T_1 and T_2 to produce this point. The solution set forms an algebraic variety, which in our trials was always of dimension 2. A primary decomposition of the ideal showed there were three components of the solution set, two of dimension 2 and one of dimension 1.

One of the 2-dimensional components was defined in part by setting one internal edge length on T_1 to infinity and one internal edge length on T_2 to 0. The mixing parameter, all split parameters in T_2, and all but 4 split parameters in T_1 were uniquely determined. Two quadratic relationships in 2 variables each held for the remaining parameters. The other 2-dimensional component is similar, with the roles of T_1 and T_2 reversed. The 1-dimensional component requires that an internal edge on each tree have length 0, but allows the mixing parameter to vary along with two edges on each tree. (See [3] for the precise results.)

It is worth highlighting that the only 2-tree mixtures matching the 1-tree distribution were of this extreme nature, with some internal edges of length 0 or infinity. If one allows these values, then there are instances of all mixture parameters being in a stochastically meaningful range. Of course formally establishing any conjectures these calculations suggest would require a detailed semi-algebraic analysis of these models.

A second question Theorem 22 might lead one to ask is if T_3 is a tree all of whose quartets comes from either T_1 or T_2, then is $V_{T_3} \subseteq V_{T_1} * V_{T_2}$? However, we have already seen an instance where this failed in Lemma 20. It would be interesting to characterize precisely when these types of containments arise.

Finally, it is not at all clear if the containment in Theorem 22 is a special phenomenon for the JC model, or if it can occur more generally for other group-based or more general phylogenetic models. Answering such questions will require an understanding of this phenomenon beyond the computational perspective.

VI. Identifiability of Continuous Parameters

Assuming tree topologies are already known, we next explore the generic identifiability of the continuous parameters in group-based mixture models. We use both rigorous arguments and computational approaches to address this issue. While standard laptop computers were sufficient for most of this work (see [3]), the more intensive computations were performed on a more powerful machine provided by Erich Kaltofen of NCSU.

Proving a model has identifiable continuous parameters requires showing that the parameterization map is one-to-one. Without any special assumptions on the map, it may be one-to-one on some region of parameter space, but not on another. Well-known to algebraic geometers, however, is that parameterization maps defined by polynomial formulas, such as those for the models we study, have the nice feature that they exhibit a *generic* behavior. More specifically, there is some $k \in \{1, 2, 3, \ldots, \infty\}$ such that for all parameter values except those in some exceptional set E, the map will be k-to-one (*cf.*, Prop. 7.16 of [17]). Crucially, the set E where the generic behavior may fail is closed and of Lebesgue measure 0 within the full parameter space (since it is a proper algebraic subvariety). In the case of complex univariate polynomials, this fact is more widely familiar: given an n-th degree polynomial $p(z)$, for almost all $\alpha \in \mathbb{C}$ the equation $p(z) = \alpha$ has n distinct roots. However, for a finite number of exceptional values of α there may be fewer distinct roots. Thus p defines a generically n-to-one map from $\mathbb{C}$ to $\mathbb{C}$.

One can computationally determine the generic behavior *with high probability* as follows: For a specific choice of parameters, calculate the cardinality of the set of all other choices of parameters with the same image. If, for many such random choices, one finds this fiber is of size k, there can be little doubt that the map is generically k-to-1. These computations can be performed exactly by computational

algebra software such as Macaulay2 [15] or Singular [16], and carefully performed repeated trials can give one high confidence. Of course, such an approach does not rigorously establish results. However, the use of random data to reliably study behavior of specific polynomial equations is not novel. For instance, Section 6 of [21] gives a different application of the idea in phylogenetics.

This approach unfortunately does not give any quantifiable meaning to the term 'high probability,' as we lack any explicit information on the set E where non-generic behavior may arise. If a non-zero multivariate polynomial vanishing on E were known, we would only need to compute that the map was k-to-one for a single point not satisfying that polynomial, and obtain a rigorous result. If we knew only the degree of such a polynomial, by the Schwartz-Zippel Theorem (*cf.*, for instance, [26]), we could produce points with arbitrarily small probability of lying in E, and use these to quantify our terminology. However, we have no such information, and thus our confidence in having determined the generic behavior is based partly on experience. In choosing points for calculations, a useful heuristic is to pick coordinates to be random rational numbers (perhaps also requiring that they be expressible using disjoint sets of primes), in hopes that the unknown polynomial equations describing E are less likely to be satisfied. Indeed, if 25 points chosen in this way all produce the same value of k, while it is possible they all lie in E, the evidence is strong that they do not.

We label statements with "Theorem*" or "Proposition*" if we are only highly confident of them through such computation. Unstarred statements are rigorously proved. Thus while we are careful to distinguish between results with rigorous proof and those depending on such calculations, we are highly confident of both.

One of the results we found computationally was a particularly surprising non-identifiability result for continuous parameters of 4-taxon tree mixtures under the JC model. Nonetheless, passing to 5-taxon trees restores identifiability.

The first main result in this section is:

Theorem 23:* For the JC model, the continuous parameters in the 2-tree mixture with parametrization ψ_{T_1,T_2} are generically identifiable for binary trees with $n \geq 4$ leaves, except in the case that $n = 4$ and $T_1 = T_2$.

An issue that will arise in our proof of Theorem* 23 and related results, concerns the maps ψ_T parametrizing V_T for the JC, K2P, and K3P models. The proof in [9] of the identifiability of numerical parameters for the general Markov model shows that identifiability of numerical parameters only holds up to permutation of states at internal nodes of the tree. Permuting the states at an internal node corresponds to permuting rows of transition matrices on edges leading out of the node, and columns of matrices on edges leading into the node. As any permutation of the rows or columns of a JC matrix that is also a JC matrix is identical to the original matrix, this implies the JC parameterization is generically one-to-one. For a generic K2P matrix, there are two orderings of the rows that have K2P form, and hence the parametrization map is generically 2^{n-2}-to-one. For a generic K3P matrix, there are four orderings of the rows that have K3P form, and

hence the parametrization map is generically 4^{n-2}-to-one. To avoid complications in statements due to these understood failures of identifiability in its strictest sense, it is more convenient to focus on the k-to-oneness of the maps ϕ_{T_1,T_2}, using equation (1) to relate results to ψ_{T_1,T_2}.

The first step toward Theorem* 23 is performing computations to establish the following.

Proposition 24:* Let $T_1 \neq T_2$ be binary trees with four leaves. Then for the JC model, the map

$$\phi_{T_1,T_2} : V_{T_1} \times V_{T_2} \times \mathbb{P}^1 \dashrightarrow V_{T_1} * V_{T_2}$$

is generically one-to-one.

Proof: [Calculation] From randomly chosen rational parameters in the domain of ψ_{T_1,T_2}, we computed a point $p \in V_{T_1} * V_{T_2}$. Solving the system of polynomial equations $\psi_{T_1,T_2}(s_1, s_2, \pi) = p$ determines the (complex) preimage of p. This preimage can be calculated using Gröbner bases, and was found to consist of a single point for the many such random choices we made. We can be therefore be highly confident that ψ_{T_1,T_2} is one-to-one, by the existence of a generic behaviour of any polynomial map. That ϕ_{T_1,T_2} is one-to-one then follows from the fact that ψ_{T_1} and ψ_{T_2} are generically one-to-one parameterizations of V_{T_1} and V_{T_2}.

Code is provided in the supplementary materials [3]. ■

Although we attempted to perform similar calculations to extend Proposition* 24 to the K2P and K3P models, these failed to terminate in 3 weeks time.

In the case of a mixture on two trees with the same topology, the possibility of interchanging the mixture components shows the map cannot be one-to-one. Generic identifiability thus corresponds to generic two-to-oneness in this case. For this type of mixture, we are able to perform computations for both the JC and Kimura models.

Proposition 25:* Let T be a binary tree with four leaves. Then for the K2P and K3P models, the map

$$\phi_{T,T} : V_T \times V_T \times \mathbb{P}^1 \dashrightarrow V_T * V_T$$

is generically two-to-one. For the JC model, the map $\phi_{T,T}$ is generically twelve-to-one.

Proof: [Calculation] The calculations which indicate this holds with high probability is similar to that for Proposition* 24.

Code is provided in the supplementary materials [3]. ■

Note that the twelve-to-oneness in the case of the JC model is not merely a mathematical anomaly relevant to complex parameter choices only. This type of non-identifiability for secant parameters can and does occur for stochastically meaningful parameters.

Example 26: Searches of parameter space give instances of 2, 4 or 8 stochastic parameter choices producing the same image in the 4-taxon 2-class JC mixture on a single tree topology. For instance, if $a^e = a_C^e = a_G^e = a_T^e$ denotes the JC parameters for one class on T, and $b^e = b_C^e = b_G^e = b_T^e$ those for the other class, and π the proportion of the first class, then

$$\pi = 0.1,$$
$$a^1 = 0.05, \ a^2 = 0.10, \ a^3 = 0.12, \ a^4 = 0.04, \ a^{12|34} = 0.01,$$
$$b^1 = 0.04, \ b^2 = 0.14, \ b^3 = 0.10, \ b^4 = 0.11, \ b^{12|34} = 0.46,$$

and 7 other choices of parameters have the same image. Up to interchanging classes, there are 4 essentially different choices. Code verifying this example, and examples showing 2 or 4 biologically relevant preimages, are included in the supplementary materials [3]. We do not know if exactly 6, 10, or 12 biologically relevant preimages can occur.

We rigorously establish the following:

Proposition 27: Let T be a binary tree with five leaves. Then for the JC, K2P, and K3P models the map

$$\phi_{T,T} : V_T \times V_T \times \mathbb{P}^1 \dashrightarrow V_T * V_T$$

is generically two-to-one.

The proof of Proposition 27 depends on a result of J. Kruskal concerning uniqueness of rank 1 tensor decompositions for 3-way arrays. As this has been exploited elsewhere [2], [7] to study identifiability of models, we give only essentials here. If M_1, M_2, M_3 are three matrices with r rows, and $\boldsymbol{\pi}$ is an r-element vector, let $\mathbf{m}_j^i$ denote row i of matrix M_j. Let

$$[\boldsymbol{\pi}; M_1, M_2, M_3] = \sum_{i=1}^{r} \pi_i \mathbf{m}_1^i \otimes \mathbf{m}_2^i \otimes \mathbf{m}_3^i.$$

The form of Kruskal's theorem most useful for our purposes is the following, from [2].

Theorem 28: (Kruskal) Let $\boldsymbol{\pi}$ be an r-element vector of non-zero numbers, and M_1, M_2, M_3 three matrices with r rows, all of whose row sums are 1. Let I_i, the *Kruskal rank* of M_i, be the largest integer such that every set of I_i rows of M_i is independent, and suppose

$$I_1 + I_2 + I_3 \geq 2r + 2.$$

Then if $[\boldsymbol{\pi}; M_1, M_2, M_3] = [\boldsymbol{\pi}'; M_1', M_2', M_3']$, there is a permutation P such that

$$\boldsymbol{\pi} = P\boldsymbol{\pi}', \quad M_1 = PM_1', \quad M_2 = PM_2', \quad M_3 = PM_3'.$$

Proof: [Proof of Proposition 27] To fix notation, let T have non-trivial splits $\{12|345, 123|45\}$ and let ρ denote the internal node on the pendant edge leading to leaf 3. Denote by ψ the natural parameterization of V_T, in terms of the entries of 4×4 Markov matrices. As there is no advantage to working in Fourier coordinates here, we use standard ones for V_T and $V_T * V_T$.

With id the identity map on $\mathbb{P}^1$, it is enough to show the parameterization map $\phi_{T,T} \circ (\psi \times \psi \times id)$ of $V_T * V_T$ is generically two-to-one.

The map $\phi_{T,T} \circ (\psi \times \psi \times id)$ can be made explicit as follows: Root the tree at ρ, and assign stochastic matrices to the edges of the tree giving conditional probabilities of state changes along those edges. If π is the mixing parameter, and $\mathbf{u} = (1/4, 1/4, 1/4, 1/4)$, then an 8-element vector

$$\boldsymbol{\pi} = (\pi\mathbf{u}, (1-\pi)\mathbf{u}),$$

gives the state distribution at the root. On the pendant edge leading to leaf i, a 8×4 matrix M_i composed of two stacked Markov matrices $M_i^{(1)}$ and $M_i^{(2)}$ of the appropriate form gives conditional probabilities, while on the internal edges there are 8×8 block-diagonal matrices $M_{12|345}$ and $M_{123|45}$ with two

4×4 blocks $M_{12|345}^{(1)}$, $M_{12|345}^{(2)}$ and $M_{123|45}^{(1)}$, $M_{123|45}^{(2)}$ of an appropriate form. Thus the superscript (1) or (2) refers to the class in the mixture.

Now the 8×16 matrix $M_{12} = M_{12|345}(M_1 \otimes^{row} M_2)$, where $\otimes^{row}$ denotes tensor products of corresponding rows, gives probabilities of observing pairs of states at leaves 1 and 2 conditioned on the state at ρ. A similar 8×16 matrix product M_{34} gives probabilities of observing pairs of states at leaves 4 and 5 conditioned on the state at ρ.

For any choice of $\boldsymbol{\pi}$ and the $M_e^{(1)}, M_e^{(2)}$, the image X under $\phi_{T,T} \circ (\psi \times \psi \times id)$ has the same entries as the 3-way array $[\boldsymbol{\pi}; M_{12}, M_3, M_{45}]$. But for generic choices of parameters, one can check that M_{12} and M_{45} have Kruskal rank 8, and M_3 has Kruskal rank ≥ 2. Indeed, one need only check that this holds for a single choice of the parameters, since then the condition, which is defined by polynomial inequalities, can fail only on a proper subvariety of the parameter space. For instance, choosing $M_1^{(1)} = M_2^{(1)}$ to be a JC matrix with off-diagonal entry 0.1, $M_1^{(2)} = M_2^{(2)} = I_4$, and $M_{12|345} = I_8$, a calculation shows M_{12} has rank 8, and hence Kruskal rank 8.

Applying Theorem 28, we get that $\boldsymbol{\pi}$, M_{12}, M_3, and M_{45} are all uniquely determined up to simultaneous permutation of rows.

However, because of the special form of the Markov matrices for the models we consider, for generic JC parameters there are exactly two orderings to the rows of M_3 so that it is two stacked blocks of the correct form, and these differ by simply interchanging the blocks. Thus we may recover $M_3^{(1)}$ and $M_3^{(2)}$, up to order. Fixing the ordering of the rows of M_3 so $M_3^{(1)}$ is on top fixes an ordering of the rows of $\boldsymbol{\pi}$, $M_{12|345}$, and $M_{123|45}$ as well. Letting a superscript of 1 denote the top 4 rows, and a superscript of 2 the bottom 4 rows of an 8-row matrix, the mixture distribution can be written as

$$\pi \left[\mathbf{u}; M_{12}^{(1)}, M_3^{(1)}, M_{45}^{(1)} \right] + (1-\pi) \left[\mathbf{u}; M_{12}^{(2)}, M_3^{(2)}, M_{45}^{(2)} \right].$$

But this weighted sum is simply the weighted sum of the two points in the image of ψ corresponding to the two classes. As ψ is known to be generically one-to-one, the parameterization of $V_T * V_T$ is two-to-one in the JC case.

As discussed following the statement of Theorem* 23, for the K2P and K3P models there are additional orderings of the rows of M_3 so that is is two stacked blocks of the correct form. Regardless of which ordering we choose, however, by arguing as in the preceding paragraph we are led to the same two points in the image of ψ_T. Thus for these models also we see the parameterization of $V_T * V_T$ by $\phi_{T,T}$ is two-to-one. ∎

Note that the use of Kruskal's theorem in this proof extends to a 2-class CFN mixture model on a 5-taxon tree, as then the Kruskal ranks of the matrices M_{12} and M_{45} are generically 4, while M_2 has Kruskal rank ≥ 2 . Although we do not focus on that model here, we record the result, as it is helps place the examples of [24] for 4-taxon trees into context.

Proposition 29: Let T be a binary tree with five leaves. Then for the CFN models the map

$$\phi_{T,T} : V_T \times V_T \times \mathbb{P}^1 \dashrightarrow V_T * V_T$$

is generically two-to-one.

Proposition* 30: For the JC 2-tree mixture model on 5-taxon binary trees, the continuous parameters are generically identifiable.

Proof: If T_1 and T_2 have no cherries in common, then all their induced quartet trees disagree. Thus applying Proposition* 24 to all 4-taxon marginalizations shows all parameters are generically identifiable.

If T_1 and T_2 have 2 cherries in common, they are identical, and Proposition 27 gives the claim.

If T_1 and T_2 have a single cherry in common, we may assume they are $T_1 = \{12|345, 123|45\}$ and $T_2 = \{12|345, 124|35\}$. Also, since the parameters are generic, we may assume the mixing parameter giving the class size for the T_1 component is $\pi \neq 1/2$. Then marginalizing to the taxa $\{1, 3, 4, 5\}$ and applying Proposition* 24 identifies the parameters on 4 edges of each of the trees, as well as the class size π for T_1.

Marginalizing to quartets involving taxa 1 and 2, and applying Proposition* 25 to them, there are 12 points in a generic fiber. However, such a generic fiber will have 6 distinct pairs of values $\{\pi, 1 - \pi\}$, and we use the value of the mixing parameter π determined above to match parameters with T_1 and T_2. ∎

The results above allow us to argue for the generic identifiability of parameters claimed in Theorem* 23.

Proof: [Proof of Theorem* 23] The $n = 4$ case is Proposition*s 24, and the $n = 5$ case is Proposition* 30.

For $n > 5$ leaves, by assuming that the parameters are generic we may also suppose the mixing parameter $\pi \neq 1/2$.

By marginalizing to 5-taxon subsets, and applying Proposition* 30, we may identify parameters on each pair of induced 5-taxon trees, but we must determine which come from which tree. If there is at least one 5-taxon subset for which T_1 and T_2 induce different subtrees, then we know the class size parameter π for T_1. Using this known value, we can determine which induced 5-taxon parameters arise from T_1 and which arise from T_2, even when the 5-taxon subtrees are topologically the same. If all 5-taxon subtrees of T_1 and T_2 agree, so $T_1 = T_2$, then we instead use the value of π_1 to collect 5-taxon subtree parameters from each copy of the tree. As the parameters for T_1 and T_2 are elements of the collection of induced parameters, we thus identify all parameters on the full trees. ∎

In closing, note that the arguments in the proof of Theorem* 23 in combination with the results of Proposition 27, rigorously prove the following result, in the case of identical tree topologies.

Theorem 31: For the JC, K2P, and K3P models, the continuous parameters in the 2-tree mixture on the same tree topology are generically identifiable for binary trees with $n \geq 5$ leaves.

REFERENCES

[1] E. S. Allman, C. Ané, and J. A. Rhodes. Identifiability of a Markovian model of molecular evolution with gamma-distributed rates. *Advances in Applied Probability*, 40:229–249, 2008. arXiv:0709.0531.

[2] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 2009.

[3] E. S. Allman, S. Petrović, J. A. Rhodes, and S. Sullivant. Supplementary material. Available at http://www4.ncsu.edu/~smsulli2/Pubs/TwoTreesWebsite/twotrees.html, 2009.

[4] E. S. Allman and J. A. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Math. Biosci.*, 186(2):113–144, 2003.

[5] E. S. Allman and J. A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comput. Biol.*, 13(5):1101–1113, 2006.

[6] E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. in Appl. Math.*, 40(2), 2008.

[7] E. S. Allman and J. A. Rhodes. The identifiability of covarion models in phylogenetics. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 6(1):76–88, 2009.

[8] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. of Class.*, 4:57–71, 1987.

[9] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

[10] D. Cox, J. Little, and D. O'Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer-Verlag, New York, second edition, 1997.

[11] J. Draisma. A tropical approach to secant dimensions. *J. Pure Appl. Algebra*, 212(2):349–363, 2008.

[12] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser, Basel, 2008.

[13] S. N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Statist.*, 21(1):355–377, 1993.

[14] W. Fulton. *Introduction to Toric Varieties*, volume 131 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1993. The William H. Roever Lectures in Geometry.

[15] D. R. Grayson and M. E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/, 2002.

[16] G.-M. Greuel, G. Pfister, and H. Schönemann. SINGULAR 3.1.0 — A computer algebra system for polynomial computations. 2009. http://www.singular.uni-kl.de.

[17] J. Harris. *Algebraic Geometry: A First Course*. Springer-Verlag, New York, 1992.

[18] M. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Systematic Zoology*, 38:310–321, 1989.

[19] M. D. Hendy and D. Penny. Spectral analysis of phylogenetic data. *J. Class.*, 10:1–20, 1993.

[20] M. D. Hendy and D. Penny. Complete families of linear invariants for some stochastic models of sequence evolution, with and without the molecular clock assumption. *J. Comp. Biol.*, 3(1):19–31, 1996.

[21] S. Hoşten, A. Khetan, and B. Sturmfels. Solving the Likelihood Equations. *Found. Comput. Math.*, 5:389–407, 2005. arXiv:math.ST/0408270.

[22] J. A. Lake. A rate independent technique for analysis of nucleic acid sequences: Evolutionary parsimony. *Mol. Bio. Evol.*, 4(2):167–191, 1987.

[23] F. A. Matsen, E. Mossel, and M. Steel. Mixed-up trees: the structure of phylogenetic mixtures. *Bull. Math. Biol.*, 70(4):1115–1139, 2008.

[24] F. A. Matsen and M. A. Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Syst. Biol.*, 56(5):767–775, 2007.

[25] E. Mossel and E. Vigoda. Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science*, 309:2207–2209, 2005.

[26] S. Rudich. Complexity theory: from Gödel to Feynman. In *Computational complexity theory*, volume 10 of *IAS/Park City Math. Ser.*, pages 5–87. Amer. Math. Soc., Providence, RI, 2004.

[27] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[28] D. Speyer and B. Sturmfels. The tropical Grassmannian. *Adv. Geom.*, 4(3):389–411, 2004.

[29] M. A. Steel and Y. X. Fu. Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model. *J. Comp. Biol.*, 2(1):39–47, 1995.

[30] B. Sturmfels. *Gröbner Bases and Convex Polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI, 1996.

[31] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12(2):204–228, 2005.

[32] L. Székely, P. L. Erdös, M. A. Steel, and D. Penny. A Fourier inversion formula for evolutionary trees. *Applied Mathematics Letters*, 6(2):13–17, 1993.

[33] L. A. Székely, M. A. Steel, and P. L. Erdös. Fourier calculus on evolutionary trees. *Adv. in Appl. Math.*, 14(2):200–210, 1993.

[34] D. Štefankovič and E. Vigoda. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions. *J. Comput. Biol.*, 14(2):156–189, 2007.